Code and Data

https://github.com/jz48/MixupExplainer

MixupExplainer: Generalizing Explanations for Graph Neural Networks with Data Augmentation

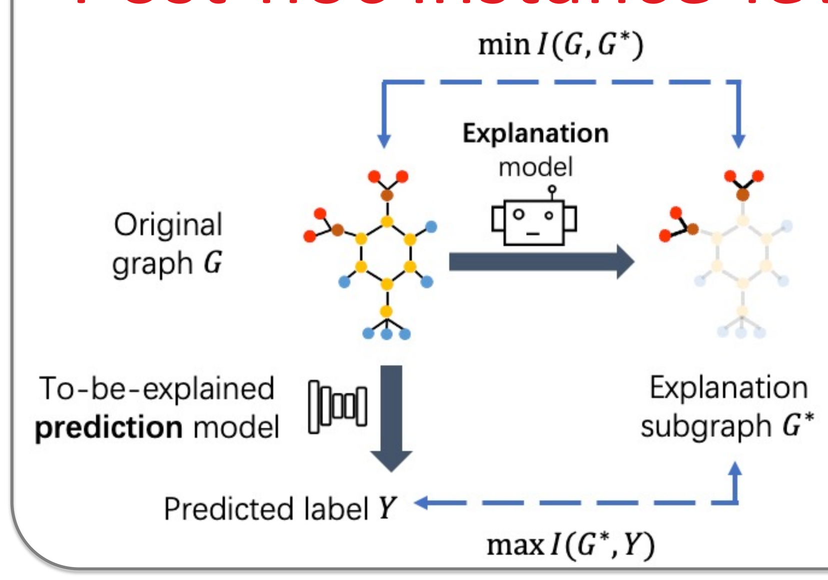
Jiaxing Zhang*, Dongsheng Luo*, Hua Wei

jz48@njit.edu, dluo@fiu.edu, hua.wei@asu.edu

Arizona State University
FLORIDA INTERNATIONAL UNIVERSITY
NJIT New Jersey Institute of Technology

Explaining GNNs

Post-hoc instance-level explanation



Discover **subgraphs** that explain the prediction of a trained GNN.

$$\arg\min_{G^*} I(G, G^*) - \alpha I(G^*, Y)$$

Graph Information Bottleneck (GIB) Objectives in a nut shell : What was right and what was wrong?

Mutual information $I(G^*, Y) = H(Y) - H(Y|G^*)$

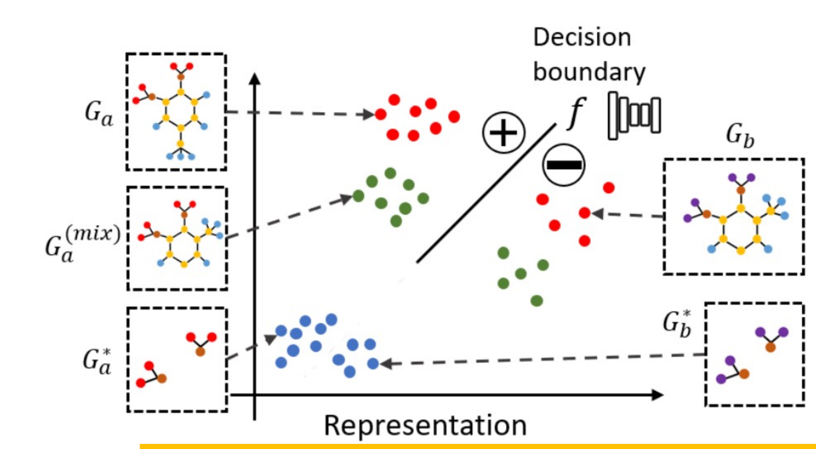
$$\arg\min_{G^*} I(G, G^*) + \alpha H(Y|G^*)$$

Intractability of $H(Y|G^*)$

$$\arg\min_{G^*} I(G, G^*) + \alpha CE(Y, f(G^*))$$

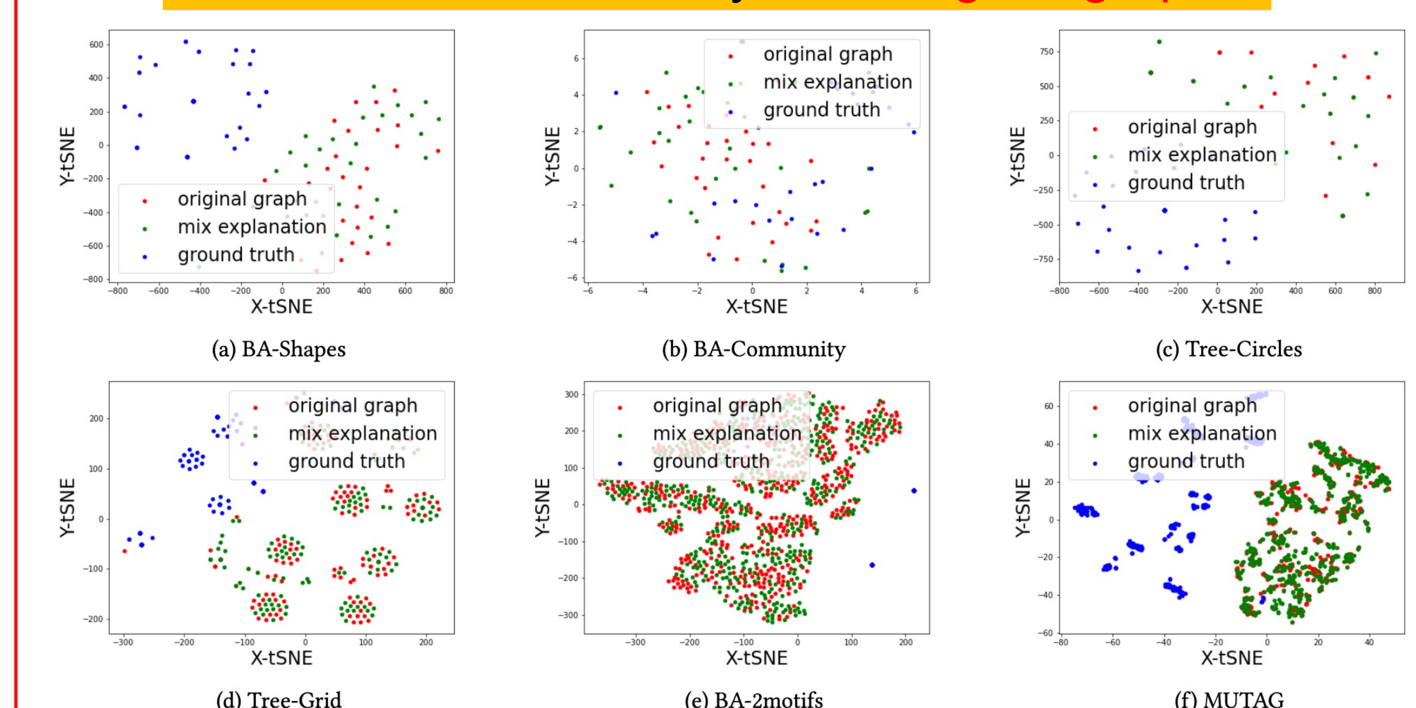
[1,2,3]

Diverging Distributions between Y and $f(G^*)$



If we visualize the output before the last layer of the model to be explained f, then plotted with t-SNE, we can find that:

Ground truth shifts away from original graph

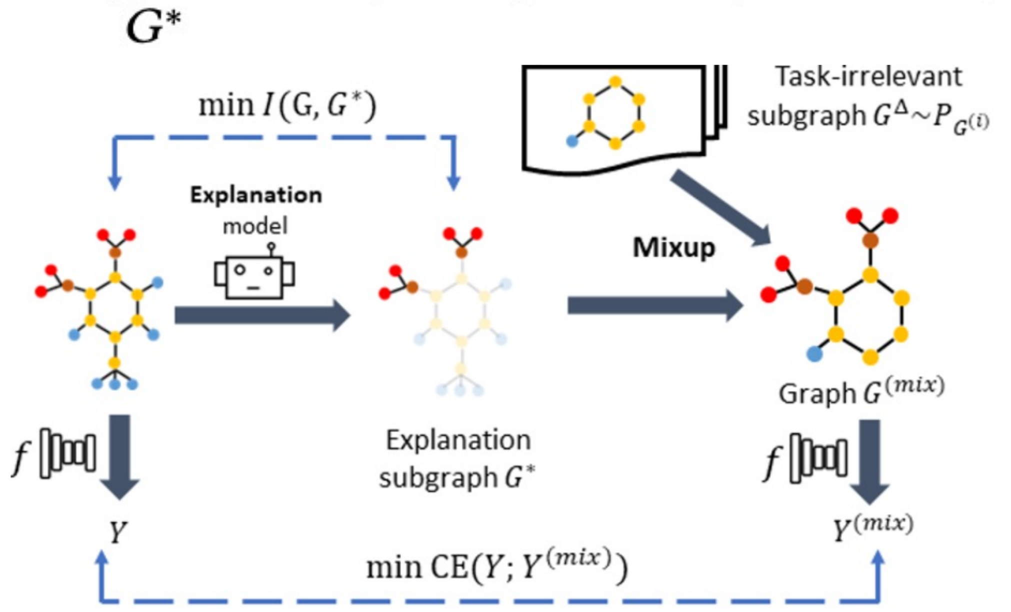


(a) BA-Shapes (b) BA-Community (c) Tree-Circles (d) Tree-Grid (e) BA-2motifs (f) MUTAG

Mix explanation aligns well with original graph

Implementation

$$\arg\min_{G^*} I(G, G^*) + \alpha \mathrm{CE}(Y, Y^{(\mathrm{mix})})$$



Intuition: An explanation on a graph from GNN is the subgraph that can mix up with any random graphs and yet does not change GNN's prediction.

Algorithm 1 Graph Mixup Algorithm

Input: Graph $G_a = (X_a, A_a)$, a set of graphs $\mathcal{G}$, the number of random connections η, explanation model g.

Output: Graph $G^{(\mathrm{mix})}$.

Randomly sample a graph $G_b = (A_b, X_b)$ from $\mathcal{G}$
Generate mask matrix $M_a = g(G_a)$
Generate mask matrix $M_b = g(G_b)$
Sample η random connections between G_a and G_b as A_c
Mixup adjacency matrix $A^{(\mathrm{mix})}$ with Eq. (10)
Mixup edge mask $M^{(\mathrm{mix})}$ with Eq. (11)
Mixup node features $X^{(\mathrm{mix})} = [X_a; X_b]$
return $G^{(\mathrm{mix})} = (X^{(\mathrm{mix})}, M^{(\mathrm{mix})} \odot A^{(\mathrm{mix})})$

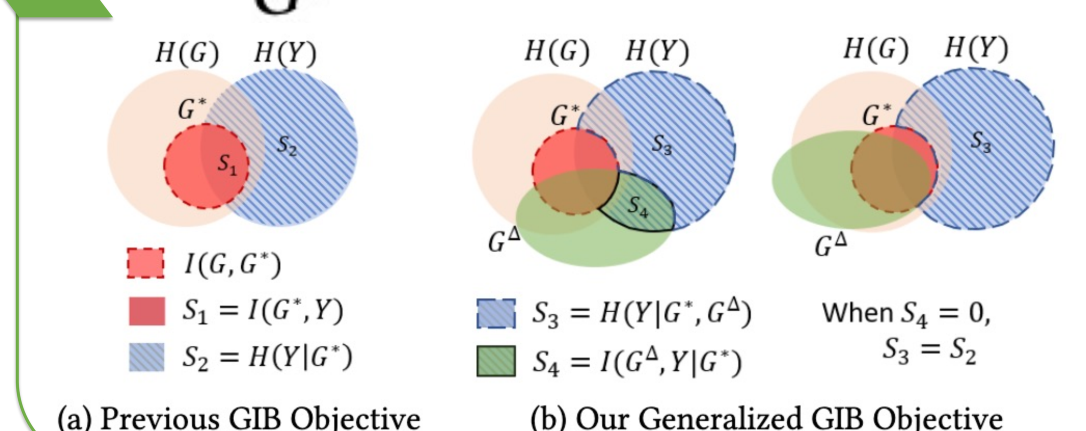
$$A^{(\mathrm{mix})} = \begin{bmatrix} A_a & A_c \\ A_c^T & A_b \end{bmatrix}$$

$$M_a^{(\mathrm{mix})} = \begin{bmatrix} \lambda M_a & M_c \\ M_c^T & A_b - \lambda M_b \end{bmatrix}$$

Nice Property: the proposed mixup approach could reduce the distance between the explanation and original graphs.

Proposed generalized GIB objective

$$\arg\min_{G^*} I(G, G^*) + \alpha H(Y|G^*, G^\Delta) \quad \text{s.t.} \quad I(G^\Delta, Y|G^*) = 0.$$



(a) Previous GIB Objective (b) Our Generalized GIB Objective

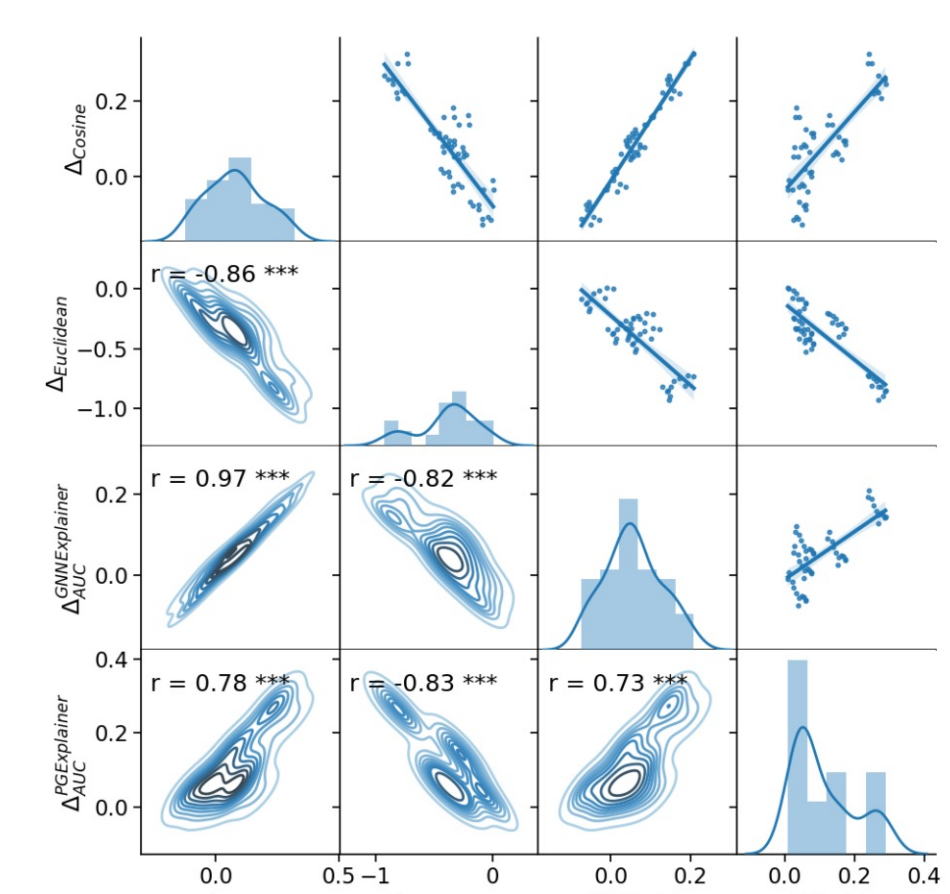
Nice property: The generalized GIB objective is equivalent to vanilla GIB

Figure 2: Illustration of GIB and our proposed new objective. (a) Previous vanilla GIB objective aims to minimize $I(G^*, Y)$ and $H(Y|G^*)$, with a smaller overlap between G^* and G. (b) Our generalized GIB objective has the same objective as vanilla GIB, with a larger lap between G and $G^* + G^\Delta$, resulting in less distribution shifting issue.

Experiments

Table 1. Explanation faithfulness in terms of AUC-ROC on edges.

	BA-Shapes	BA-Community	Tree-Circles	Tree-Grid	BA-2motifs	MUTAG
GRAD	0.882	0.750	0.905	0.612	0.717	0.783
ATT	0.815	0.739	0.824	0.667	0.667	0.765
SubgraphX	0.548	0.473	0.617	0.516	0.610	0.529
MetaGNN	0.851	0.688	0.523	0.628	0.500	0.680
RG-Explainer	0.985	0.919	0.787	0.927	0.657	0.873
GNNExplainer	$0.884_{\pm 0.002}$	$0.682_{\pm 0.004}$	$0.683_{\pm 0.009}$	$0.379_{\pm 0.001}$	$0.660_{\pm 0.006}$	$0.539_{\pm 0.002}$
+ MixUp	$0.890_{\pm 0.004}$	$0.788_{\pm 0.006}$	$0.690_{\pm 0.014}$	$0.501_{\pm 0.003}$	$0.869_{\pm 0.004}$	$0.612_{\pm 0.043}$
(improvement)	0.60%	15.5%	1.02%	32.2%	31.7%	13.5%
PGExplainer	$0.999_{\pm 0.001}$	$0.829_{\pm 0.040}$	$0.762_{\pm 0.014}$	$0.679_{\pm 0.008}$	$0.679_{\pm 0.043}$	$0.843_{\pm 0.084}$
+ MixUp	$0.999_{\pm 0.001}$	$0.955_{\pm 0.017}$	$0.774_{\pm 0.004}$	$0.712_{\pm 0.000}$	$0.920_{\pm 0.031}$	$0.871_{\pm 0.079}$
(improvement)	0.00%	15.2%	1.57%	4.86%	35.5%	3.32%



All these four improvements strongly correlated to each other with statistical significance: The improvements achieved by MixupExplainer in explanations accuracy own to the successful alleviation of the distribution shifting issue.

Conclusion

- Be careful if you are using GNNExplainer or PGExplainer! You might encounter distribution shifting issue between $f(G)$ and $f(G^*)$.
- An explanation of a GNN's prediction on an original graph is the subgraph that can mix up with any random graphs and does not change GNN's prediction.

References

[1]. Ying, et al, GNNexplainer: Generating explanations for graph neural networks. NeurIPS 2020.
[2]. Luo, et al., Parameterized explainer for graph neural network. NeurIPS 2020.
[3]. Miao et al., Interpretable and generalizable graph learning via stochastic attention mechanism. ICML 2022.

The work was partially supported by NSF award #2153311. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.